Prof. Dr. Matthias O. Franz

Übung 1: Explorative Analyse und Vorverarbeitung

Praktikum Maschinelles Lernen

1. Explorative Analyse des Datensatzes "Boston Housing"

Boston Housing ist ein berühmter Datensatz zur Evaluierung von Regressionsalgorithmen. Er enthält 506 Einträge mit jeweils 13 Variablen. Ziel ist es, den Hauspreis ('tgt') aus den anderen Variablen vorherzusagen. Der Download dieses Datensatzes in einen Pandas-DataFrame wird folgendermaßen durchgeführt:

```
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/housing/
    housing.data'
cols  =
    ['CRIM','ZN','INDUS','CHAS','NOX','RM','AGE','DIS','RAD','TAX','PTRATIO','B',
    'LSTAT','TGT']
boston = pd.read_csv(url, sep=' ', skipinitialspace=True, header=None,
    names=cols, index_col=False)
```

Wichtig für diese Übung ist eine grundlegende Vertrautheit mit den Python-Paketen Numpy und Pandas. Die Abgabe der Aufgabe erfolgt als fertiges IPython-Notebook mit Kommentaren in Markdown.

Aufgaben:

a. Führen Sie für diesen Datensatz eine explorative Analyse wie in der Vorlesung gezeigt mithilfe eines IPython-Notebooks und den Paketen Pandas und Numpy durch. b. Beantworten Sie anhand der Darstellung der Streumatrix folgende Fragen: Welche der Variablen sind kategorisch? Welche der Variablen eignen sich gut zur Vorhersage des Hauspreises und warum? Welche dieser Variablen sind miteinander korreliert? Welche sind daher Kandidaten, die man evtl. weglassen könnte? (Beantwortung bitte als Markup in Notebook eintragen) c. Die Dokumentation der eingesetzten Systemkonfiguration und Paketversionen erfolgt durch das Skript 'version_information' von R. H. Johanson. Installation in älteren IPython/ Python-Versionen erfolgt durch Aufruf in IPython oder innerhalb des Notebooks von

```
%install_ext http://raw.github.com/jrjohansson/version_information/master/
    version_information.py
```

In neueren IPython/Python-Versionen wird das Paket über den Paketmanager `pip` oder `pip3` installiert:

```
pip version_information
```

Im Notebook-Header muss das Paket importiert werden über

```
%reload_ext version_information
```

Danach kann die Information über die Systemkonfiguration durch

```
%version_information
```

dargestellt werden. Sollen zusätzlich Versionsinformationen über die eingesetzten Pakete dargestellt werden, verwendet man (hier z.B. Numpy und Pandas)

```
%version_information numpy, pandas
```

Diese Zeilen sollten immer am Ende des Notebooks aufgerufen werden, um ein Mindestmaß an Reproduzierbarkeit sicherzustellen.

2. Datenvorverarbeitung mit Pandas: Food Nutrient Database

Diese Aufgabe befasst sich mit einer deutlich umfangreicheren Datenbank des *U.S. Department of Agriculture*, aufbereitet im Format JSON von A. Williams, zum Thema Nährstoffgehalt von Nahrungsmitteln. Sie enthält 6636 Einträge für Nahrungsmittel, alle in Form von JSON-Records, wie z.B.:

```
{
  "id": 21441,
  "description": "KENTUCKY FRIED CHICKEN, Fried Chicken, EXTRA CRISPY, Wing,
    meat and skin with breading",
  "tags": ["KFC"],
  "manufacturer": "Kentucky Fried Chicken",
  "group": "Fast Foods",
  "portions": [
    {
      "amount": 1,
      "unit": "wing, with skin",
      "grams": 68.0
    },

    ...
  ],
  "nutrients": [
    {
      "value": 20.8,
      "units": "g",
      "description": "Protein",
      "group": "Composition"
    },

    ...
  ]
}
```

Ziel der Analyse in dieser Übung ist es, eine explorative Analyse des Gehalts des Spurenelementes Zink in den verschiedenen Nahrungsmitteln zu durchzuführen. Notwendig dafür sind etwas aufwändigere, aber für die Datenanlyse typische Manipulationen mit Pandas sowie der

Einsatz zusätzlicher Python-Standardbibliotheken zum Download und der Verarbeitung von Zip- und JSON-Dateien.

Aufgaben:

a. Laden Sie die Datenbank als zip-File aus Moodle herunter und lesen Sie dieses File direkt in ein neues Notebook ein. Die bisher verwendete Pandas-Methode `read_csv()` funktioniert für JSON-Files leider nicht. Das heruntergeladene File wird stattdessen mithilfe des Pythonmoduls `zipfile` entpackt und dem Python-Befehl `open()` eingelesen. Die Umwandlung des JSON-Formates in ein geeignetes Python-Format erfolgt mit einem weiteren Modul der Python-Standardlibrary, `json`, hier mithilfe der Funktion `json.load()`. Lesen Sie dazu die zugehörigen, auf dem Web bzw. Stackoverflow verfügbaren Anleitungen.

b. Die Datenbank steht nun in Form einer Liste aus 6636 Python-Dictionaries zu Verfügung. Jedes Dictionary enthält Angaben zu einem Nahrungsmittel. Greifen Sie sich ein beliebiges Nahrungsmittel heraus und lassen sich die Namen der Einträge mit der Methode `dict.keys()` anzeigen. Einer der Einträge enthält die enthaltenen Nährstoffe (`'nutrients'`), ebenfalls als Dictionary. Lassen Sie sich wiederum einen beliebigen Eintrag der Nährstoffliste anzeigen. Es sollte auffallen, dass manche Feldnamen doppelt vorkommen.

Teile dieser hierarchischen Struktur sollen nun in eine einheitliche Tabelle umgewandelt werden, um eine explorative Analyse durchführen zu können.

Vorgehensweise:

- Kopieren Sie zunächst die Felder `'description','group','id','manufacturer'` in einen eigenen DataFrame `info`, sowie alle Nährstofflisten in ein Array von DataFrames, wobei Sie an jeden DataFrame die entsprechende ID des Nahrungsmittels als eigene Spalte anhängen.
- Dieses Array wird mithilfe der Funktion `pandas.concat()` zu einem großen DataFrame `nutrients` (389355 Einträge) vereinigt.
- Entfernen Sie alle Duplikate aus diesem DataFrame.
- Bevor beide DataFrames vereinigt werden können, gibt es noch ein Problem: beide enthalten Felder mit dem Namen `'description'` und `'group'` (s.o.). Benennen Sie diese daher mithilfe von `DataFrame.rename()` in eindeutige Namen um.
- Vereinigen Sie beide DataFrames mit `pandas.merge(nutrients, info, on='id', how='outer')` anhand der Nahrungsmittel-ID.

Überprüfen Sie das Ergebnis jeder Manipulation mit `DataFrame.head()`.

c. Nun sind die Daten bereit für die Untersuchung auf das Spurenelement Zink (Feldname: `'Zinc, Zn'`). Lesen Sie dazu alle Tabelleneinträge mithilfe einer geeigneten Indizierung in einen DataFrame aus, der nur Einträge zum Nährstoff Zink enthält. Daraus wählen Sie wiederum die Spalte mit dem Zinkgehalt in mg (`'value'`) aus und stellen dafür ein Histogramm und eine Liste deskriptiver Statistiken dar. Finden Sie in Ihrer Tabelle Edamer (`'Cheese, edam'`). Hat Edamer einen überdurchschnittlichen Zinkgehalt? Haben mehr als

75% aller Nahrungsmittel einen kleineren Zinkgehalt? Welches Nahrungsmittel hat den maximalen Zinkgehalt?